Overview of Matching Dates for WIC Clinics

Here is a general summary of what we found:

Number total clinics	Matched By ID	Matched By Name	Matched By Address	Matched but Nonmatching Zipcodes	Date Changed Manually
465	267	269	173	37	23

Opening Dates Found	Remaining NA Opening Dates	NA Dates due to No Match Found	NA Dates due to Bad Match (Review)	Matches with Dates Changed from NA to 1993-01-01
425	40	33	7	212

I will give a brief overview of the process.

First, I edited the datasets a little. I added two “normalizers” to better compare names and addresses. For both of these, I basically stripped the strings and got rid of common abbreviations, e.g. for names I got rid of phrases like “wic” and “mobile.” For street addresses, I replaced “north” with “n” and god rid of street abbreviations like “Lane” or “Ln.” I then sorted alphabetically to make each one unique.

The reason for this is that some names are almost the same, but not quite. For example, there might be a wic clinic in the old dataset called Abott Wic Center, but in the new dataset, is just called Abott or Abott Wic. Or, an address may be 123 Elmo Lane in one dataset, but in the other is 123 Elmo or 123 Elmo Ln.

Example of how I changed the addresses:

address	f_addr
5201 Harry Hines Blvd	0125aehhinrrsy

I also fixed up the site id formatting for the 2024 dataset and checked that the site ids are unique. I then created variables that created various combinations of siteid, zipcode, name, address, and county for future matching: [siteid + zipcode], [normalized name + zipcode + county], or [normalized address + zipcode + county].

There were two main tasks to complete:

1. In the old dataset, deciding if I should substitute in 1993-01-01 for an NA opening date or leave it as NA
2. Matching the old clinics to the new clinics

I grouped the dates in the old dataset into three main categories:

1. Both had NA values
2. Date open was NA but date close had a value
3. Date open had a value but date close was NA

When fixing dates, I put dates into a new column called fixed_date_open. This is what I decided the opening date for that clinic was.

Depending on which of the above categories the clinic fell into, I treated it slightly differently but the main "problem clinics" were those with an NA date open but a date close existed. For these clinics, I had a few main rules:

- If the clinic is the first of its kind and reopening(s) happened within 31 days
  - This clinic is assumed to be the original and we assign its opening date 1993-01-01
- If the clinic did not reopen within 31 days
  - If there were no known opening dates for any clinic of its kind, treat this as the original and assign it an opening date of 1993-01-01
  - If all clinics of its kind opened after this one closed, treat it as the original clinic and assign an opening date of 1993-01-01
- If there are clinics of its kind that opened earlier
  - This clinic is not an original therefore we leave its date open as NA
- If there are no other clinics of its kind
  - Treat it as the only clinic and assign it an opening date of 1993-01-01

Note: I did not change any NA date close values at all.

For example (the top two tables are the same rows, I just couldn't fit them in the same line):

siteid	name	zip	COUNTY	name_zip_county
8704	jefferson wic clinic	75657	marion	eeffjnors_75657_marion
3917	jefferson wic clinic	75657	marion	eeffjnors_75657_marion

date_open	date_close	fixed_date_open
NA	1998-10-31	1993-01-01
1998-11-01	NA	1998-11-01

Clearly, the siteids do not match, but these clinics are the same. You can also see that the second clinic reopened within 31 days of the first clinic. The below row is the matching row in the 2024 dataset, where I chose the earlier fixed_date_open and flagged how the matches were made:

name	by_id	by_name	by_addr	fixed_date_open
jefferson wic clinic	TRUE	TRUE	FALSE	1993-01-01

Below, you can see an example of where I do not take the earlier date open. Because the clinic reopened after 31 days, I take the opening date of the second clinic.

siteid	name	zip	COUNTY	name_zip_county
5702	commerce clinic	75428	hunt	cceemmor_75428_hunt
7665	commerce clinic	75428	hunt	cceemmor_75428_hunt

date_open	date_close	fixed_date_open
NA	1999-03-31	1993-01-01
1999-07-01	NA	1999-07-01

In the 2024 dataset, the matching row is:

name	by_id	by_name	by_addr	fixed_date_open
commerce wic	FALSE	TRUE	FALSE	1999-07-01

Sam and I initially got 21 different date opens (before I went in and changed dates manually for clincis that did not match) for non NA date opens and most of the reasons were because I was choosing date opens based on if a clinic had reopened within 31 days or not.

As mentioned, I then went in and manually changed the opening dates for clinics that were still NA and I found 23 more opening dates.